# DAT 375 Project One Template

# Data Analysis Process Job Aid

### Who should use this job aid?

The intended audience for this job aid are data analysts.

### Introduction

This job aid will support new data analysts with analyzing a data model using a standardized approach. There are four major steps to evaluate: Type of analysis, defining parameters and collecting data, tool selection, and validation. Each of these steps are fundamental for data analysis and it should be understood that every analysis may need a judgement call on adaptation depending on the project assigned.

### Section 1: Type of analysis

First we must identify and address the problem. The Miami Police Department (MPD) has identified a link between an increase in crimes with an increase in storms. They want to be able to anticipate possible timeframes for future crimes.

Knowing the problem we now need to identify the proper analysis to execute. The keyword in the problem is anticipate, this means they want to be able to predict future outcomes. Knowing that the objective is to predict the analysis method will be regression analysis which is an inferential quantitative method. (Bhatia, 2019)

### Section 2: Define Parameters and collect data

With the problem defined and the analysis method determined, we must now define our relevant parameters and identify what data to collect. The first attribute we need to measure for our observations is the storm activity. We need to record the crime activity that occurred. We also need to identify where the storm and crime location to identify the correlation between them.

Location is optional and should be reviewed for accuracy to see if it is statistically significant or not in the regression model. The two primary attributes to review will be storm activity and crime activity.

### Section 3: Tool Selection

There are three tools that can be recommended. First a database to record observations, secondly, a statistical calculation tool, and finally a visualization tool. For this project, MySQL will be the database tool to record observations and query data. For the statistical calculation tool and visualization tool, we will use R, a programming tool with excellent statistical and visualization capabilities. (Ng, 2021)

The first analysis we will calculate is a logistic regression model to identify the correlation between storms and crimes. Further analysis, such as a classification tree, can correlate which storms correlate to which crimes.

### Section 4: Validation

As requested in the initial documentation the script to identify the records that do have a direct correlation between storms and crimes is:

“SELECT \* FROM dat375.stormcrimes

WHERE CrimeEventID <>0

AND StormEventID <>0;“

However, for logistic regression modelling, this would create a survival bias in the dataset and would calculate incorrectly. For logistic regression modelling the entire data set would need to be used. (Morgenstern, 2018) It is highly recommended that in any data analysis that you review data for bias before reporting.

The selected parameters are sufficient for providing a predictive logistic regression model. However for a more accurate modelling it would be expected that every instance of crime, regardless of storm, is recorded. And that every instance of a storm, regardless of crime, is also recorded. A standardization of recording behavior would be recommended to ensure no bias is inherent in the data itself.

References

Bhatia, M. (2019, July 12). Your Guide to Qualitative and Quantitative Data Analysis Methods - Atlan: Humans of Data. Retrieved January 19, 2021, from https://humansofdata.atlan.com/2018/09/qualitative-quantitative-data-analysis-methods/

Morgenstern, J. (2018, April 7). Survival bias. Retrieved January 23, 2021, from https://first10em.com/ebm/survival-bias/

Ng, A. (2021, January 20). Top 30 Big Data Tools for Data Analysis in 2021. Retrieved January 23, 2021, from https://www.octoparse.com/blog/top-30-big-data-tools-for-data-analysis-in-2021